Aurally Aided Visual Search Performance Comparing Virtual Audio Systems

Larsen, Camilla Horne; Lauritsen, David Skødt; Larsen, Jacob Junker; Pilgaard, Marc; Madsen, Jacob Boesen; Stenholt, Rasmus

Published in:
137th International AES convention

Publication date:
2014

Document Version
Publisher's PDF, also known as Version of record

Audio Engineering Society

Convention Paper

Presented at the 137th Convention
2014 October 9–12 Los Angeles, USA

Aurally Aided Visual Search Performance Comparing Virtual Audio Systems

Camilla H. Larsen, David S. Lauritsen, Jacob J. Larsen, Marc Pilgaard, Jacob B. Madsen & Rasmus Stenholt

Aalborg University, Aalborg, 9210, Denmark

Correspondence should be addressed to Marc Pilgaard (mpilga10@student.aau.dk)

ABSTRACT

Due to increased computational power, reproducing binaural hearing in real-time applications, through usage of head-related transfer functions (HRTFs), is now possible. This paper addresses the differences in aurally-aided visual search performance between a HRTF enhanced audio system (3D) and an amplitude panning audio system (panning) in a virtual environment. We present a performance study involving 33 participants locating aurally-aided visual targets placed at fixed positions, under different audio conditions. A varying amount of visual distractors were present, represented as black circles with white dots. The results indicate that 3D audio yields faster search latencies than panning audio, especially with larger amounts of distractors. The applications of this research could fit virtual environments such as video games or virtual simulations.

1. INTRODUCTION

Spatial audio, audio that is emitted by a source with a physical position, is an important feature in virtual environments as it helps users orient themselves, and can provide 360° aural awareness, independently of sight. Applications that require precise and realistic spatial orientation often places the user in the center of the environment, rendering everything relative to the user himself. This emulates a virtual representation of the user in the virtual environment, and makes use of the user's innate hearing perspective. Applications, ranging from games, movies and music, to military equipment and training simulators, often utilize spatial audio, as it has become the industry standard. Previously, the possibility of more closely simulating binaural hearing in real-time was constrained by hardware performance, as the average home computer did not posses the processing power related to the task. Even though typical hardware can handle the processing task today, still only few applications utilize this technology. Most often, a simple stereo amplitude panning method is used for spatial sound rendering [4].

Humans hear binaurally, which makes us able to perceive the spatial position of sound, due to the arrival time difference of the sound at each ear, as well as

subtle alterations in the characteristics of a sound's frequency spectrum. This alteration is due to the shape of the pinna, head and body, as well as the reflective- and absorptive properties of clothes, skin and hair. The common method of synthesizing these properties is through head related transfer functions (HRTFs). These are functions that are based on the frequency responses of sound for each ear, when placed at different positions around the head. The method of measuring these functions, is by placing a microphone within the ears of a representative head model, which is usually calculated from a mean of many. A sound is then played at discrete positions all around the head, and the difference in the frequency response from the original sounds is derived. These are dependent on models such as head- and body shadowing, as well as the reflection and absorption of sound, based on the head model. These frequency alterations can then be applied to any audio signal, based on the spatial position at which they are played.

Previous studies have shown that the utilization of HRTFs is suitable for audio localization, in both real- and virtual scenarios [1, 7, 8, 10–14]. Prior to this study, we did an audio exclusive localization experiment, the results of which suggest that binaurally simulated sound (3D sound) improves localization performance, compared to amplitude-panned sound (panning sound) [6]. This study tries to determine if there is also a significant difference between the spatial sound rendering methods in visual search. It has previously been shown that audio aids visual search [1, 2, 5, 7, 8, 10–14], which makes performance differences relevant. Many modern, interactive applications rely heavily on visual search, making the potential improvement worth researching.

The comparison between 3D sound and panning sound is in our experiment exclusive to virtual environments. What differentiates this experiment from our previous experiment [6] is the inclusion of visual stimuli Our hypothesis is this: There is a significant difference in search latencies in aurally aided visual search, between using a panning- and a 3D audio system. Audio systems that make use of HRTFs are often more computationally heavy compared to panning systems. If search latencies are faster when using 3D systems, the results can help one in determining whether to sacrifice a bit of system performance in exchange for reduced acquisition time.

2. RELATED WORK

Gröhn, Tapio & Savioja [7] conducted a localization experiment in a virtual room. They took interest in comparing three different modalities: HRTF (non-individualized, through headphones), direct reproduction (loudspeakers) and vector based amplitude panning (loudspeakers) [16]. The experiment included visual stimuli. The tasks were to determine position, indicated by pointing with a discrete resolution of one degree, and distance of a perceived pink noise sound source. They found the non-individualized HRTF reproduction to yield the least accurate precision.

Due to the unavailability of individualized HRTFs in more practical applications, Wenzel, Kistler & Wightman [22] tried to determine if the precision of non-individualized HRTFs are sufficient. The participants were therefore tasked with localizing the azimuth and elevation of wideband noise pulses, using headphones. These results were then compared with a similar test, conducted in free field, using loudspeakers. 12 of the 16 subjects performed with sufficient accuracy in the comparison. Of the remaining four, only two showed poor elevation accuracy exclusively with the virtual sources, while the last two showed consistently poor elevation accuracy across both tests. Wenzel et al. did however see a significant increase in front-back and up-down confusion with the virtual sources. Their HRTFs were recorded from 10 subjects at a resolution of 144 source positions. [23]

Perrott et al. [13] looked at the significance of spatial audio cues in target detection, with distractors. They found a significant reduction in localization time with aural cues over silence, and a particular reduction in instances with a large number of distractors. The tests were conducted using a setup of 21 speakers and the room was isolated with acoustic foam to come as close to free field as possible. Audio pulses between 800 and 9000 Hz were used in this experiment.

McIntire et al. [8] looked at search performance in a dynamic, virtual environment, comparing 3D audio cues (Non-individualized HRTFs, headphones) and no audio. The test included distractors, using short bursts of wideband white noise. The goal of their

study was to build upon previous studies that have shown a significant increase in search performance in static environments when using 3D audio, by adding movement. The results of the paper suggest that 3D audio reduces search time, both in static and dynamic instances. Furthermore, to their own surprise, they saw a considerable initial advantage to using 3D audio in the very first samples of the test, yet no consistent improvement by adaption, compared to the no audio.

Setting out to test the efficiency of audio, visual and audio-visual cue combinations in a virtual environment, Gröhn, Lokki & Takala [5] exposed their test participants to a maze-like scenario in which they had to navigate from gate to gate to reach the end. These gates were either highlighted with audio, a visual element or both. The audio cue consisted of wideband pink noise bursts, presented in free field. The results suggest that the audio-visual combination yields the best results, that visual comes in next and that the audio cues were the least effective.

Bolia, D'Angelo & McKinley [1] tested the significance of spatial audio for target acquisition performance, including both axis (+/- 180 degrees in azimuth, +90 to -70 degrees in elevation) with visual distractors present. The audio conditions were: No audio, spatial audio (266 speakers, free field) and virtual spatial audio (Non-individualized HRTF, headphones). The results suggest a reduction in target acquisition delay by a factor of six with virtual spatial sound compared to having no audio cue at all, and that the virtual spatial sound makes no significant reduction in accuracy and target acquisition time compared to their spatial audio setup.

Simpson, Iyer & Brungart [18] look at the effects of distractors in a visual search scenario. To test this, they task their subjects with identifying the characteristics of a visual target in a field of visual distractors. The target would sometimes emit a spatial audio cue, and at other times, the cue would be accompanied by one or several audio distractors at different locations. Spatial audio cues, in the form of continuous wideband noise was given through an array of 277 loudspeakers. Simpson et al. found that it was possible for their subjects to extract spatial information from up to three simultaneous sound sources, but that search performance suffered greatly when presented with four or more simultaneous sound sources. They also found that with 15 audio distractors, it took twice as long for the subjects to locate the visual target as when there were no audio distractors. This suggests that the subjects were unable to ignore the distractors.

Perrott et al. [12] divided their work, with visual search performance with auditory cues, into three experiments. Different to most other experiments presented in this section, Perrott et al. used very few speakers for spatial sound. The target loudspeaker and visual cue were presented at the end of a pivoting boom arm. The first experiment was conducted with the target at a fixed elevation (eye level), whereas the second experiment used both vertical and horizontal planes. Perrott et al. saw a substantial reduction in target localization latency when a 10 Hz clicking sound was presented in the same position as the visual target cue. Furthermore, they found a significant decrease in localization latency when the target was located 10 degrees from the initial line of gaze. In the third experiment, Perrot et al. focused on head and eye movements when a participant was trying to locate a sound source, observing that concurrent head and eye movement are common in audio-based search. In over half of the trials, participants would shut their eyes in an attempt to orient themselves towards the sound source. This supports the hypothesis that the auditory spatial channels role in regulating visual gaze is significant.

Serving as previous study to this one, Larsen et al. [6] focused on the differences in localization performance between a 3D audio system (non-individualized HRTFs, headphones) against a panning system (amplitude panning, headphones). The test was conducted in an audio exclusive, virtual environment and a sonar sound was used as an audio cue. The experiment included three tests, each pertaining to a different aspect of localization performance: Speed, precision and navigation. The experiment proved a significant difference in favor of 3D audio on all fronts.

In relation to the articles covered, yet unique in its choice of factorial sound systems, this study focuses on the comparison between a volumetric panning system, and a non-individualized HRTF audio system in an exclusively virtual scenario. The test in-

cludes visual distractors (based on the distractors used by Bolia, D'Angelo & McKinley [1]), as these are important to avoid pop-out effects, but no audio distractors. The test makes use of bursts of pink noise, as broadband noise represents the frequency spectrum well. This study separates itself from the prior study by focusing on visual search, instead of audio-exclusive search.

3. EXPERIMENT DESIGN

In this experiment, the participant had to locate a visual object within a virtual environment with changing audio- and visual conditions for each trial. The participant's task was to locate a visual target among a set of visual distractors. A mouse was used to orient the player's view in the virtual environment. By manipulating the view, the audio output was rendered dependent on the direction to the target's position relative the player's view. The dots per inch of the mouse was fixed between all participants. The experiment consisted of two independent variables: Audio condition and amount of distractors. The audio condition consisted of three levels: No audio, panning audio and 3D audio. The amount of distractors consisted of eight levels: 0, 2, 4, 8, 16, 32, 64 or 128. The amount of distractors was determined by the assumption that search latency is linearly dependent on the amount of distractors, as seen in other experiments [1, 3]. The exponential increase in the amount of distractors is to cover a broader range of values, because we believe that deviations from the linear increase will not occur through small increments.

The test participant was seated in front of a laptop with a mouse, and was required to wear a pair of headphones. The test conductor began the experiment and the participant was told to follow the commands on the screen. The participant was told that the goal was to locate the target as fast as possible. Before the experiment began, the participant was given time in a training session within the virtual environment, in which the participant could become familiar with the interface and task. Every participant was required to go through the training session at least once. This training session used the same visual- and auditory stimuli as the real test. When the participant felt he understood the assignment, he could begin the experiment.

Upon beginning the experiment, a countdown, counting from three to zero, appeared. When the countdown reached zero it was the participant's task to locate the target. When the participant had located the target, he was required to aim at the target with an on-screen crosshair and perform a left mouse click within the stimuli's bounds. If the participant hit the target, the system took control of the virtual camera and panned it to its starting position. The countdown appeared again, and the next trial began. If the participant did not hit the target, the system would not recognize it as a successful hit and the participant did not progress, and was not informed of his miss. Aiming was restricted from -89° to 89° in the vertical axis, but the horizontal axis was unrestricted. Each participant went through 144 different trials, hence 48 trials for each audio condition. The order of audio rendering conditions followed a 3x3 Latin Square model. The experiment was conducted as a within subject experiment.

The auditory stimuli used was 700 ms bursts of pink noise at a fixed audible level, with a silence period of 700 ms. The pink noise stimuli also had 100 ms of linear fade, both in and out, leaving 500 ms at full intensity.

The visual stimuli used was a black circle with white dots in the middle. The black circle's visual angle was 4.7° of the virtual camera view. The target object was a black circle with an odd amount (one or three) of white dots inside, while the distractors had an even amount (two or four). The stimuli were also randomly rotated at 90° steps. See Figure 1 for an illustration of the visual stimuli. The design of the visual stimuli was similar to the experiment conducted by Bolia et al. In their research, they used LEDs that were grouped together as visual stimuli. The target stimuli had an even amount of LEDs and the distractors had an odd amount [1].

To ensure conscious search, it is important that the target does not simply "pop out", as this would make the number of distractors completely redundant. The reason for this is, that a drastic change from target to distractor could elicit a preattentive response. Treisman [19] found that line orientation had a big impact on early visual processing, but that a particular positional arrangement of lines do not. Furthermore, color and simple shape properties, such as curving, are important, and will cause

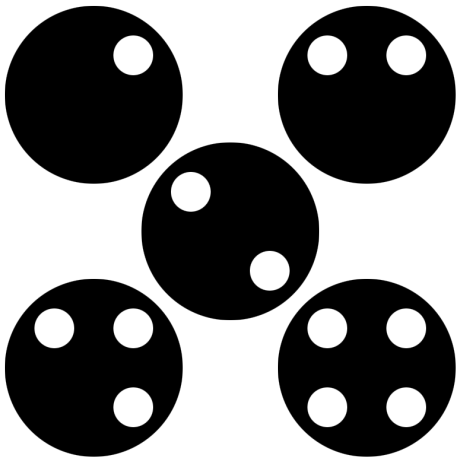

Fig. 1: The visual stimuli used within the experiment. The ones with an even amount of white dots are distractors, while those with an odd amount are targets.

the pop-out effect. Early visual processing reacts to individual features, not conjunctions of features. In other words, if the features that are reacted to (For example color, brightness and line orientation) in the early visual stages are kept constant, they can be combined or arranged positionally without popping out. As an example, Treisman suggests that the letters T and L do not pop out from one another, because the important features match (One vertical line, one horizontal line). Based on this, we deemed our stimuli, which was based on Bolia et al.'s stimuli, to not get caught by preattentive search.

The positions of both the target and distractors were based on the vertices of a icosahedron, with 6 m in diameter. It was subdivided three times, resulting in 162 available points. The distractors and the target were placed at random discrete positions, each at one of the 162 available positions. There was only one target per trial. The amount of distractors for each trial was chosen through fixed randomization, assuring balance across all conditions. All stimuli were facing the participant's avatar at all time. The virtual settings was inside a box, the sides of which were dimly lit by yellow light. This was to provide participants with a sense of orientation. See figure 2 for a visual example of how one set might look. The participant's avatar was in the center of the icosahedron, with a distance of 3 m to every point. A meter within the virtual environment can not directly be translated into an actual meter, though the audio rendering engines approximate attenuation models based on the distance in meters.

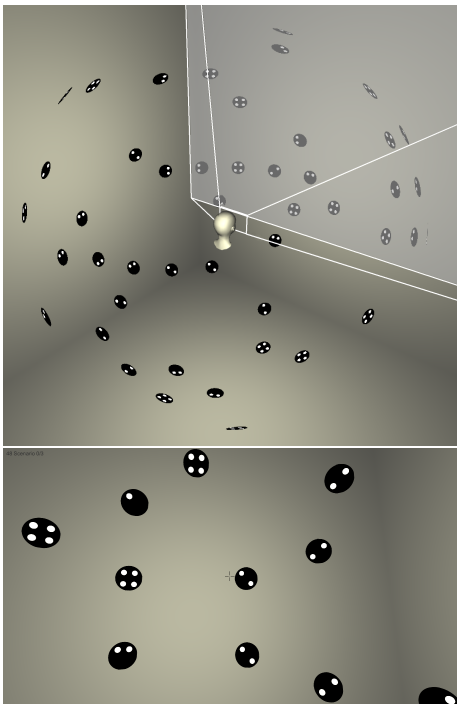

Fig. 2: Two depictions from the virtual environment. The top image shows a the setup of the environment with the virtual character in the center and 129 instances (note that the backfacing visuals cannot be seen) of visual stimuli distributed on the vertices of the icosahedron. The highlighted area represents the visual field. The bottom image is an example of how the screen could appear to a subject.

3.1. Experiment setup

The virtual environment ran at a 1366 x 768 resolution with a field of view (FOV) at 90° horizontally and 60° vertically, with a frame rate above 60 frames per second. The virtual environment ran on two laptops simultaneously and independently. The first was an ASUS KS53SV laptop with a Mobile Intel HM65 Express Chipset and the second was an ASUS K55A, using a Intel Chief River Chipset HM76. Both chipsets utilize Intel High Definition Audio technology. All audio altering effects were disabled, and equalization options were kept stan-

dard on both machines. For user input, a Logitech MX 518 and a Logitech G400 was used, with a dpi set at 1600. For audio playback, a pair of Sennheiser 360 G4ME headsets were used.

For audio rendering, Diesel Power Mobile (DPM), developed by AM3D[1], rendered the 3D audio. Diesel Power Mobile used a set of non-individualized HRTFs based on a study performed by Møller et al. [9]. The HRTFs is based on the averages of multiple real-life users with a resolution of 22.5°and is further upsampled to 5.625°. Interpolation of the four closest HRTFs is used. The specific rendering approach is confidential to AM3D. FMOD[2], developed by Firelight Technologies, rendered the panning audio, and makes use of interaural intensity difference. The experimental software was developed in Unity3D[3], where FMOD is a part of the default installation. A plugin was developed for implementing DPM into Unity which consists of a C# wrapper of the DPM library. The C# wrapper communicates positions, orientations and audio signals to the DPM library, which returns a mixed filtered stereo signal. The returned stereo signal is then directly injected into the pipeline between FMOD and the computer speakers. Note that FMOD does not process the signal returned from the DPM library. A more thorough explanation of the implementation can be found in the paper from Larsen et al. [6].

4. RESULTS AND DISCUSSION

In the experiment, 31 males and 2 females participated with ages ranging from 20 to 30 years ($\bar{x}$ = 22.5). It took each participant an average of 14.5 minutes to complete the 144 trials. All participants reported normal or corrected to normal hearing and sight. All participants were university students. 4752 samples were recorded, 1584 for each audio condition, 198 samples for each amount of distractors. The confidence coefficient was set at 95%.

Due to a non-homogeneous variance in search time (Bartlett Test of Homogeneity of Variances, $p < 0.001$), most tests conducted were of a non-parametric nature, as transforming the data would still maintain its non-homogeneity.

[1] See `http://www.am3d.com/` for company information
[2] See `http://fmod.org` for fuinformationrther product
[3] See `http://unity3d.com/` for further product information

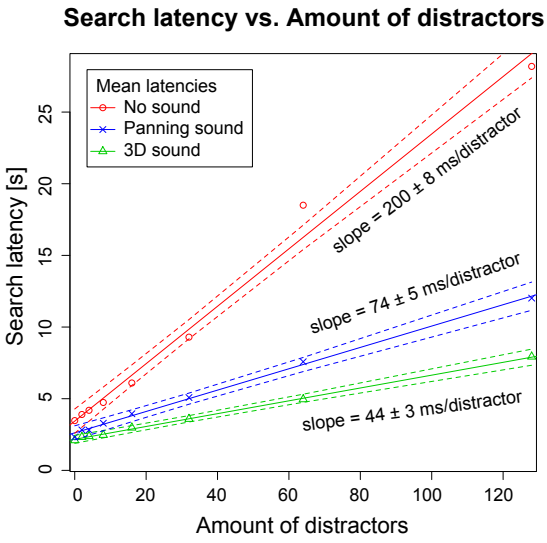

Fig. 3: Mean latencies through the different audio conditions when increasing the distractor amount. The solid lines are tendencies of the mean latencies, and the dotted lines represent 95% confidence intervals.

A significant difference in search latencies was identified between the different audio conditions independent of distractor amount (Friedman test and multiple comparison of treatments, $p < 0.001$) being $\bar{x}$ = 9.798 seconds for the no audio condition, $\bar{x}$ = 3.588 seconds for the 3D audio condition and $\bar{x}$ = 4.985 seconds for the panning audio condition.

There was also a significant difference in search latencies with an increased amount of distractors (Friedman test and multiple comparison of treatments, $p < 0.001$), see Figure 3. There was a significant difference between all distractor amounts, with an exception between two and four distractors (Friedman test and multiple comparison of treatments, $p > 0.05$) $\bar{x}$ = 3.014 seconds and $\bar{x}$ = 3.128 seconds respectively. There was also a significant difference in the interaction between the auditory conditions and the amount of distractors (Friedman test and multiple comparison of treatments, $p < 0.001$). Table 1 presents all the mean latencies and standard errors of the means for every possible condition. Figure 4 illustrates the different combinations of audio conditions and distractor amounts and their

significance level to each other. The linear increase in search latencies at increased distractor amounts suggests that the search task was not affected by preattentive search.

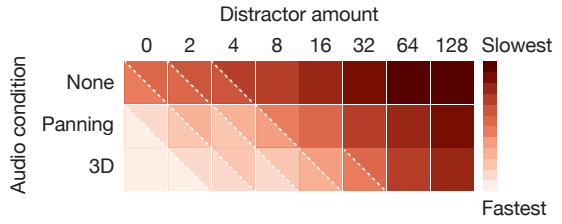

Fig. 4: The significance levels between combinations of audio conditions and amounts of distractors. White represents the combinations with the fastest search latency, red represents the slowest. Squares of different color are of a significant difference. Squares with two colors are of no significant difference to both identically colored whole squares.

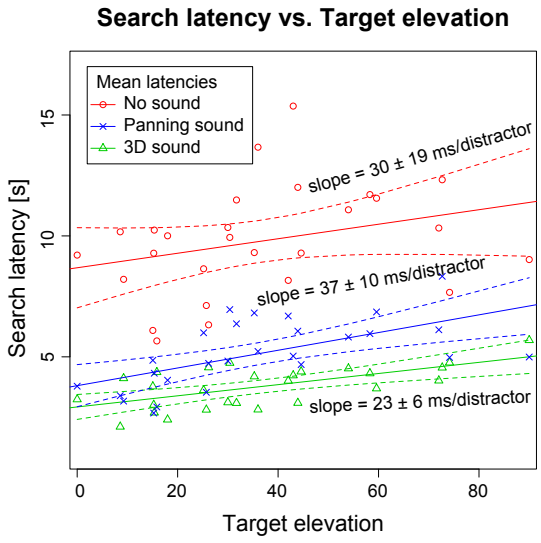

Fig. 5: Mean latencies through the different audio conditions when increasing elevation from start position. The solid lines are the tendencies of the mean latencies, and the dotted lines represent 95% confidence interval.

Different starting positions for the target also created a significant difference (Friedman test and multiple comparison of treatments, $p < 0.001$). If the target stimuli was placed at a vertical position exceeding 30° in the vertical plane relative the participant's initial look direction (making it not visible from the start) the participant would have to search through the vertical plane in order to reveal the target. These conditions increased the search latencies significantly (Friedman test and multiple comparison of treatments, $p < 0.05$) with 0.802 seconds for the 3D audio condition and 2.194 seconds for the panning audio condition, see Figure 5. Because positions were randomly selected between a set of 162 fixed positions, the data was unbalanced, meaning there was an unequal quantity of data entries for the different conditions, and so down sampling of the 4752 trials was necessary. The entries were discarded at random. 4752 entries were recorded originally, but down sampling took this number down to 4506, leaving 1502 samples per audio condition. All entries were treated as absolute, because we were predominantly interested in the differences in search latencies between positions inside and outside of the initial field of view, vertically. The entries were divided into two groups. The first group contained all the starting positions between 0° and 30° of elevation, positions which would fit into the participant's FOV without searching the vertical plane. The second contained all starting positions between 30° and 90° of elevation, positions that required the participant to search in the vertical plane. On Figure 5 it can be seen that search time is linearly dependent on the elevation of the target, and that the line representing panning is steeper than that of 3D, which is most likely due to the missing vertical audio cues in panning audio. For the no sound condition, this seems to be random. This could be further investigated by looking into search patterns.

The accumulated time each participant had the target within their FOV, when fully visible, was also recorded. There was a significant difference (Friedman test and multiple comparison of treatments, $p < 0.001$) in time, with mean values of $\bar{x} = 1.932$ seconds, $\bar{x} = 1.603$ seconds and $\bar{x} = 1.451$ seconds respectively for the no audio, panning audio and 3D audio conditions. It is interesting to see that latencies were improved, even when the target was within the visual field, where one might assume visual perception to be dominant. This could be due to adaptive behavior, as the participant relies on au-

Table 1: This table represents the means and standard errors of the means in combinations of audio condition and amount of distractors.

		Amount of distractors							
		0	**2**	**4**	**8**	**16**	**32**	**64**	**128**
Mono	$\bar{x}$ **(s)**	3.472	3.900	4.192	4.738	6.098	9.296	18.502	28.187
	SEM	0.357	0.480	0.437	0.557	0.824	1.341	3.856	5.190
Panning	$\bar{x}$ **(s)**	2.094	2.321	2.381	2.467	2.975	3.583	4.953	7.930
	SEM	0.208	0.227	0.202	0.251	0.323	0.507	0.830	1.913
3D	$\bar{x}$ **(s)**	2.325	2.822	2.812	3.289	3.949	5.082	7.579	12.019
	SEM	0.234	0.477	0.334	0.417	0.624	1.079	1.333	3.358

dio for an early orientation cue. Typical orientation through binaural hearing is helped by moving the head and listening, potentially explaining the result. Head movement was observed in earlier studies [12] and McIntire et al. [8] also observed a significant difference in search latencies within the FOV.

The visual stimuli could take two forms: A black circle with either a single white dot or three white dots. The data indicates that there was a significant difference in search latencies between using the different stimuli (Friedman test and multiple comparison of treatments, $p = 0.003$) with mean latencies of $\bar{x} = 7.0$ seconds for the stimuli with three white dots and $\bar{x} = 5.347$ seconds for the stimuli with a single white dot. A few participants also reported that it was more difficult to locate the stimuli with three white dots. It was randomly chosen which visual stimuli to use and so the data collection was unbalanced and down sampling of the 4752 trials was necessary. The entries were discarded at random. 4752 samples were down sampled to 4602 samples, leaving 2301 occurrences for each visual stimuli. The significant difference in latencies between the two different visual target stimuli suggests that they were not equally susceptible to visual attention. Based on the works of Pomerantz and Cragin [15] we believe this is due to differences in emerging features. Both distractors and the three dot target stimuli contain more than two black dots, therefore emerging features such as proximity and orientations between the two or more points is present. The single dot stimuli can not have these features, except its positional feature from its single dot. This makes it stand out and so becomes a Gestalt, leading it to emerge from the field of distractors. This can lead to faster acquisition times. These findings are of interest, though they do not have a great impact on our experiments results due to the within-subject design. These results suggest that visual stimuli needs to be considered carefully, for it not to have impact on the results of an experiment. If testing active visual search, it is important that the stimuli does not elicit the participants' attention in preattentive search through emerging features.

The results indicate that the use of spatial audio aids in visual search task, compared to using no sound. This supports the previous findings which suggests that audio aids in visual search tasks [1, 2, 8, 11–13]. Also the results indicate that using 3D audio increases search performance, compared to using panning audio. The results also indicate that an increased amount of visual distractors increases search latencies, which also supports various previous experiments [1, 14, 17]. Furthermore, the results support that target detection takes longer for targets located vertically outside of neutrally elevated field of view [7, 8, 22]. Lastly, even when the target are within the FOV, 3D audio still yields significantly faster search latencies than panning audio. This suggests that spatial audio still aids in visual search tasks when the target is within the visual field of view.

5. CONCLUSION

In this study we identified the difference in visual search performance in a virtual environment, using

either 3D audio or panning audio. The results show that using 3D audio compared to panning audio decreases search latencies significantly (by 28%), confirming previous studies [8]. The results also show that with increasing amount of visual distractors, 3D audio reduces search latencies compared to panning audio. This suggests that the visual distractors affect search performance, across all audio conditions. Another finding was that search latencies were significantly decreased, by 9.5%, for 3D audio compared to panning audio, even with the visual stimuli being within the field of view, suggesting that auditory stimuli is used as an aid to visual search even within the field of view. The two visual stimuli gave significantly different search times, where the single-dot stimuli elicited faster search than the three-dot stimuli, which implies that the fewer features a visual stimuli have, the easier it is to find.

The results from this study further adds to the observation that 3D audio yields better visual search performance than panning audio, which we find of interest to software developers and researchers interesting in exploring the effects and uses of 3D audio.

6. FUTURE WORK

The results of this experiment can be applied to simple virtual environments where stimuli are few and non-complex. These results may not be very applicable for environments that include larger quantities of both visual- and auditory stimuli, such as modern computer games and simulators. It is therefore of relevance to perform further research on this topic within more complex environments.

7. REFERENCES

[1] Robert S. Bolia, William R. D'Angelo, and Richard L. McKinley. Aurally aided visual search in three-dimensional space. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 41(4):664–669, December 1999.

[2] Patrick Flanagan, Ken I. McAnally, Russell L. Martin, James W. Meehan, and Simon R. Oldfield. Aurally and visually guided visual search in a virtual environment. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 40(3):461–468, September 1998.

[3] Scott M. Galster, Robert S. Bolia, and Raja Parasuraman. Effects of information automation and decision-aiding cueing on action implementation in a visual search task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 46(3):438–442, September 2002.

[4] Mark Grimshaw, editor. *Game Sound Technology and Player Interaction: Concepts and Developments*. IGI Global, September 2010.

[5] Matti Gröhn, Tapio Lokki, and Tapio Takala. Comparison of auditory, visual, and audiovisual navigation in a 3D space. *ACM Trans. Appl. Percept.*, 2(4):564–570, October 2005.

[6] Camilla H. Larsen, David S. Lauritsen, Jacob J. Larsen, Marc Pilgaard, and Jacob B. Madsen. Differences in human audio localization performance between a HRTF- and a non-HRTF audio system. 2013.

[7] T. Lokki M Gröhn. Using binaural hearing for localization in multimodal virtual environments. 2001.

[8] John P. McIntire, Paul R. Havig, Scott N. J. Watamaniuk, and Robert H. Gilkey. Visual search performance with 3-d auditory cues: Effects of motion, target location, and practice. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 52(1):41–53, February 2010.

[9] H. Møller, C. B. Jensen, D. Hammershøi, and M. F. Sørensen. Using a typical human subject for binaural recording. Audio Engineering Society, May 1996.

[10] W. Todd Nelson, Lawrence J. Hettinger, James A. Cunningham, Bart J. Brickman, Michael W. Haas, and Richard L. McKinley. Effects of localized auditory information on visual target detection performance using a helmet-mounted display. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 40(3):452–460, September 1998.

[11] David R. Perrott, John Cisneros, Richard L. Mckinley, and William R. D'Angelo. Aurally aided visual search under virtual and free-field

listening conditions. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 38(4):702–715, December 1996.

[12] David R. Perrott, Kourosh Saberi, Kathleen Brown, and Thomas Z. Strybel. Auditory psychomotor coordination and visual search performance. *Perception & Psychophysics*, 48(3):214–226, May 1990.

[13] David R. Perrott, Toktam Sadralodabai, Kourosh Saberi, and Thomas Z. Strybel. Aurally aided visual search in the central visual field: Effects of visual load and visual enhancement of the target. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 33(4):389–400, August 1991.

[14] Andrea C. Pierno, Andrea Caria, Scott Glover, and Umberto Castiello. Effects of increasing visual load on aurally and visually guided target acquisition in a virtual environment. *Applied Ergonomics*, 36(3):335–343, May 2005.

[15] James Pomerantz and Anna Cragin. Emergent features and feature combination. *Oxford Handbook of Perceptual Organization*, To be published.

[16] Ville Pulkki. Virtual sound source positioning using vector base amplitude panning. *Journal of the Audio Engineering Society*, 45(6):456–466, June 1997.

[17] Monique Radeau and Paul Bertelson. Auditory-visual interaction and the timing of inputs. *Psychological Research*, 49(1):17–22, June 1987.

[18] Brian Simpson, Nandini Iyer, and Douglas S. Brungart. Aurally aided visual search with multiple audio cues. volume 2010, Washington, D.C, USA, June 2010. ICAD.

[19] Anne Treisman. Features and objects in visual processing. *Sci. Am.*, 255(5):114–125, November 1986.

[20] Erik Van der Burg, N. L, Adelbert W. Bronkhorst, and Jan Theeuwes. Pip and pop: Nonspatial auditory signals improve spatial visual search. *Journal of Experimental Psychology: Human Perception and Performance*, 34(5):1053–1065, 2008.

[21] Qinqin Wang, Patrick Cavanagh, and Marc Green. Familiarity and pop-out in visual search. *Perception & Psychophysics*, 56(5):495–500, September 1994.

[22] Elizabeth M. Wenzel, Marianne Arruda, Doris J. Kistler, and Frederic L. Wightman. Localization using nonindividualized head-related transfer functions. *The Journal of the Acoustical Society of America*, 94(1):111–123, 1993.

[23] F L Wightman and D J Kistler. Headphone simulation of free-field listening. i: Stimulus synthesis. *The Journal of the Acoustical Society of America*, 85(2):858–867, February 1989. PMID: 2926000.